L3: Entry 5 of 5 File: USPT Nov 4, 2003

DOCUMENT-IDENTIFIER: US 6643641 B1
TITLE: Web search engine with graphic snapshots

Abstract Text (1):
A search engine manages the indexing of web page contents and accepts user selection criteria to find and report hits that meet the search criteria. The inventive search engine has an associated crawler function wherein display images of the web pages are rendered and stored as snapshots, preferably when the pages are indexed. The search engine reports search results by composing an html page with links to the corresponding page hits and containing snapshot reduced size graphic images showing the web pages as they appeared when fetched and stored as snapshots.

Brief Summary Text (14):
The search engine operator can use various methods to find or select web page addresses that will be loaded and analyzed or indexed in building the database. The methods may be chosen to expand or to limit the number of web pages that the search engine will access. As a result, the results of searches vary among the different search engines.

Brief Summary Text (17):
Examples of search engines include Hotbot, AltaVista, Yahoo, NorthernLight, Excite, etc. In addition, there are some search engine portals that run the same user query through a plurality of other search engines. The search engine comprises a processor that maintains a web page which the user loads by aiming his browser at the search engine URL (e.g., Excite's URL is http://www.excite.com/). The received page (namely the processed version of the html source code that is displayed) typically includes one or more Common Gateway Interface (CGI) boxes or similar form processing means by which a user who wishes to make a search enters one or more letter strings as search criteria. Boolean combinations of two or more strings often can be included or will be implied if not stated. The criteria typically are construed met if the specified words or phrases are found anywhere in the html source code of the target pages when last indexed. This includes portions that are not displayed (e.g., meta-tags and comments). The criteria can specify attributes other than the presence anywhere of a certain text string. This may be helpful, for example, to limit search results to finding files of a certain type (e.g., with URLs linking to a certain file extension type to find a certain kind of media). The criteria can also bracket out files in a selected date window.

Brief Summary Text (21):
The typical search engine reports more to the search than the URLs of the indexed pages that meet the searcher's selection criteria. The URLs themselves, which are formatted as hypertext links in the search report, sometimes provide information as to whether or not a search hit is pertinent to the user's desires. For example the domain name associated with the page may identify an owner known to be in a pertinent business, or on the contrary may show that the search result is plainly not relevant to the search. The search engine typically also stores and includes in the search report listing one or two of the first lines of the web page that is referenced, which frequently includes a title that may be helpful to show quickly whether the selected page is of interest. The search listing also may show the date at which the web page was last updated or the date that it was indexed.

Brief Summary Text (23):
It would be advantageous if the presentation of search results could be supplemented to more effectively assist a user running a search to quickly and meaningfully separate the pertinent and irrelevant results. However, such a capability will only be useful if it can be accomplished without unduly adding processing time and storage requirements to the steps

involved in preparing database information for search and in presenting the results to the user.

Brief Summary Text (25):
It is an object of the invention to provide an abbreviated representation of searchable data files, in particular Internet/Intranet/Extranet html data pages, which represents their text and linked graphics in a visual snapshot form to supplement representations such as introductory text passages and URL addresses. It is a further object to collect and process the necessary information before conducting searches and to store a relatively small graphic file in association with the search database for representing each potential hit. The respective graphics file is reported to the user when a search results in a hit on the file, namely by inserting a hyperlink to the stored file in the search report sent to the user as the search results.

Brief Summary Text (29):
These and other objects are accomplished by the improved search engine of the invention, for managing user search and selection of data files stored at distributed systems coupled at network addresses. In particular the search engine is effective to improve searching of hypertext web pages on the Internet. The search engine has an associated web crawler operable to address and load successive web pages, and to index text data associated with the successive web pages. In this manner the search engine obtains parameter information such as words appearing in documents, word proximity and other information that can be used to distinguish at least groups of the web pages from one another when conducting a search. The web crawler stores the parameter information in a manner that cross references the paramater information with the associated web addresses or URLs of the web pages. The search engine accepts user-submitted search criteria and conducts a search or the parameter information to select the associated addresses of web pages that met all or part of the search criteria. The results can potentially be ranked, subdivided into categories and similarly handled according to known search engine operation. According to an inventive aspect, in conjunction with obtaining the parameter information for at least a subset of the web pages subject to search, the crawler renders a display image of the web page that is being indexed, and processes the image to provide a reduced size graphic image file corresponding to a static visual presentation of each of the indexed web pages. This graphic image file preferably is stored in a compressed graphic file format such as GIF, JPG, or a similar file, the file address or URL of which is stored and cross referenced to the criteria in the database that identifies the corresponding web page. When a search is conducted and results in a hit on a web page, its graphic snapshot is linked to the search results reported to the user. In a preferred embodiment, acceptance of the user search criteria and reporting of the results are handled by html page exchange communications between the search engine and the user. The search engine is accessed by the user and provides a form page having CGI boxes or the like for accepting text and/or other selections from the user. The search engine conducts a search which identifies one or more hits that are reported to the user by sending an html search results page. The search results page is composed by the search engine as a function of the search results and may contain no hits or a number of hits. Each of the hits is identified in the search results by the graphic snapshot, and preferably also by text information that reflects the content of the web page hit. Preferably, the search results page is composed to include a hypertext link to the URL address where the graphic snapshot file has been stored by the web-crawler/database/search-engine processes, for example by an IMG SRC=[path.backslash.filename] command inserted in html source code. As a result, the image file is loaded by the user's browser when processing the search results page, which generally occurs after the display of text has been accomplished.

Brief Summary Text (30):
As a result, the search results appearing on the user's browser include links to the web pages that were found to meet the criteria (hits), and also a snapshot graphic image of the way that the web page appeared when rendered at the time of indexing.

Brief Summary Text (32):
According to an inventive aspect, the graphic image file that is produced is not necessarily identical to the appearance of the page when ultimately loaded by the user after a search. In addition to the fact that the web page may have changed since it was rendered into the graphic file, the rendering is accomplished according to a predetermined display configuration of the crawler when rendered. Nevertheless, the graphic is a useful and very quick means for a user to sift through search results and determine immediately whether or not at least some of the hits

bear further investigation.

Drawing Description Text (5):
FIG. 3 is a block diagram illustrating operation of the invention in connection with executing and reporting the results of searches.

Detailed Description Text (2):
According to the invention as generally shown in FIGS. 1-3, the reporting of search results by a search engine 20, is improved and facilitated by offering each searcher or user 30 a visual representation 35 of the web pages found to meet the user's search criteria submitted to the search engine. The invention is particularly applicable to an Internet search engine but can also be applied to other networks 50 where the search engine 20 is available for managing user search and selection of web pages or similar files, stored at distributed systems 52 coupled to the network. The web pages, which may be considered data files, are found at addresses to which the search engine can link to load the data files, for example being accessible using URL addressing of the pages as hypertext markup language (html), file transfer protocol (ftp), telnet or other such file types. The data files may have embedded links to other data file or to graphics or other media files. The search engine 20 of the invention accepts user queries that characterize files of interest, searches for the files and reports to each such user the results of the search including network addresses of the files found to at least partly meet the query, enabling the user to link directly to the files, and also a snapshot of how the file will appear according to the most recent rendering performed by the crawler of the search engine.

Detailed Description Text (6):
A block diagram showing an improved Internet search engine 20 according to the invention, for managing user search and selection web pages stored at distributed systems 52 coupled at network addresses to the Internet 50 or the like, is shown generally in FIG. 1. FIG. 2 illustrates a succession of method steps and/or programmed operations of the system for building and adding to or updating a database 62 of searchable information. FIG. 3 illustrates a method and apparatus for conducting searches by accepting user queries 54, conducting searches of the database 62 and reporting search results in the form of a composed search report 80 containing visual representations or snapshots 35 that depict a presentation of how the selected pages would have appeared according to a default display configuration at the time they were accessed by the crawler 60.

Detailed Description Text (14):
According to an inventive aspect, the crawler 60 that is operable to receive the web pages and to extract the parameter information from them, generates a file 72 of graphic image data corresponding to an appearance of each of the web pages, which is stored, preferably as a reduced-size and compressed image data file 75, in association with the database data respecting the page. When search results are reported to the user (FIG. 3), the search engine reports the associated URL addresses 82 of web pages that met the search criteria in a conventional manner, preferably inserting a hypertext link to each identified page into an html page reported to the user, optionally a short description or excerpt, and also inserts into the report page the graphic image snapshot file by inserting into the source of the report page a link to the stored compressed graphic image file 75. The user's browser displays the search results in conventional form, namely by showing a selectable hyperlink to the addresses and optionally a description or excerpt, and displays a snapshot of how the identified page is likely to appear if or when it is loaded by the user's browser, should the user point and click to the link to invoke the URL of the page hit.

Detailed Description Text (15):
The search portal 78 that performs the search by reference to the database 62 in storage media 64, reports the search by composing a web page containing the search results, assembling the search report using hypertext markup language. The search report contains headers and information identifying the portal and perhaps contains advertising. The search report also lists the hits that resulted from the search. More particularly, the search engine inserts (in list or table form) a text string showing the URL address of each web page hit (i.e., the pages found to meet the user criteria) together with a hypertext linkage to that URL (e.g., an "href=" statement), causing the user's browser to show a link that can be invoked (pointed and clicked) to load the page at the stated address. Such a report is conventional in an html source search report. It typically also has a description or excerpt and may be arranged in a

pyramid or hierarchy of categories. According to the foregoing inventive aspect, the search engine also inserts the URL address of the graphic file that has been processed by a further process identified in FIG. 2 as Web Agent B 95, to contain a snapshot reduced/compressed graphic 35 representing the page hit.

Detailed Description Text (16):
The link to the compressed rendered graphic file can be made, for example, by use of a IMG SRC=<domain>/<path><filename> command in the html source. The graphic can be associated with a hypertext link to the hit page URL as well as linking using an HREF=<URL of hit page> command as mentioned above. As a result, the user's browser when displaying the search results also displays the graphic snapshot image, as shown in FIG. 3.

Detailed Description Text (18):
Referring to FIG. 2., the search engine includes or is associated with web crawler 60, which is an engine that conducts web page addressing, loading and analyzing, and stores representative data in a storage device 64 containing a database 62. The stored representative data characterizes the web pages that the crawler loads and that are analyzed for content by process 68. Of the main activities to be effected by the search engine system (i.e., by the crawler and the search processor), preparation of database 62 allows a search to be conducted more quickly by reference to the processed database information gleaned from the field of possibly-selected files, than would be possible if the search engine attempted to load and analyze the entire universe of files after the user had submitted query 54 (FIG. 3), namely while the user was awaiting search results.

Detailed Description Text (28):
The search/reporting steps of the browser, generally shown in FIG. 3, include accepting search criteria 54 from user 30, for example using a CGI script technique in which the user enters selections including text strings, literal strings of plural terms, additional encoded aspects such as media types, date windows or limits, countries of origin, etc. The user may also select Boolean relationships (AND, OR, NOT, XOR). The search portal may require commands or may permit selection using point-and-click steps. The search engine compares the search criteria to the pre-prepared database of information gleaned from the web pages it has loaded and analyzed from the field. The results are reported to the user by preparing and formatting an html source reporting page into which hyperlinks are entered that name and point to the addresses of the files that were found to meet the criteria. Often the report includes other information such as the date the page was last updated before it was indexed, and a few lines of introductory text from the page, which provide a hint to assist the user in determining without loading the page whether the page is likely to be relevant to the search. If the user finds a link that appears to be pertinent, the user selects and engages the hyperlink. This causes the browser to load the html source found at the URL address shown in the search report, and any referenced files and links therein. However, the page may have changed between the time that the indexing was accomplished and may have totally different content than it had when indexed. The page may no longer exist. In those cases, the search fails except to advise the user that the page formerly held information that might have been of interest.

Detailed Description Text (29):
Deliberate as well as inadvertant "search engine corruption" sometimes occurs. It may be crucial for marketing or other purposes for a web site to be found in user searches on search engines, and it can be lucrative or otherwise beneficial for a web site operator if his/her site is ranked high in the search results for particular terms. Thus, a great number of website operators have ways to misrepresent the content of their pages. Keywords intended to cause the page to be selected and to rate highly in particular categories can be included and may or may not be displayed. Misleading text can be placed in miniscule font at the bottom of a page or misleading text can be hidden by making it the same color as the background on which it appears. Text can also be placed in "ALT" descriptions of images and graphics, thereby indexed by the crawler but not seen by the user. A particular term can be included one or many times to improve rankings, by one of the foregoing techniques, or by overloading keywords in "META" tags included in web pages and not displayed. Another technique is to temporarily post a page to be textually indexed by the crawler/search engine and then to replace its content after it has been indexed, or similarly, meta-refreshing the web page so as to redirect the user to another page address. According to an aspect of the present invention, the user can visually distinguish pages having undesired content and not waste time on them. Search engine corruption using the aforementioned techniques to provide misleading text is averted due to the visual

nature of the present invention.

Detailed Description Text (31):
The snapshots 35 can be contained in formatted image files (e.g., GIF, JPG, etc.). The snapshot image files, or URL addresses pointing to the image files, preferably are stored in the database 62 that also contains the URL addresses of the indexed pages. In reporting search results, the search engine 78 inserts a link 82 aiming to the snapshot image file 35 into the html search results page 80. The search results appear on the users browser 84 as a link to selected pages with an associated snapshot of the page when indexed, as shown in FIG. 3.

Detailed Description Text (52):
The search engine reports search results to the user that entered the search criteria, by composing an html source page and transmitting it to the user. This html report page may identify no hits or a long list of hits, depending on the search results. In composing the report page, the search engine typically shows the search criteria used, and displays indicia summarizing or similarly identifying each web page hit. For example, the search report can identify hits by the URL of the originating web page. Preferably a short text selection such as the first few lines of text is shown. The html coded report page prepared by the search engine includes an associated hyperlink to the URL of each hit. The URL can be shown in plain text and provided with an associated hypertext link (href=[URL]). The user reviews the URLs, sample text or other information and activates the hyperlink of a selected web page identified in the results, thereby loading the web page presently found at the address of the originating page when processed by the crawler robots.

Detailed Description Text (64):
The two general functions associated with preparing the database of information which is then subject to search and reporting, are the functions of retrieving all webpage data (performed by Web Agent A), and generating a "snapshot" file from the data (performed by Web Agent B). It is found that these functions can operate concurrently with or apart from the search engine processor or processors that search the database of information and return results to the requesting user. The preferred embodiment, however, is to perform all processing in regards to rendering, resizing, and compressing the snapshot prior to being accessible to surfers on the web. A cycle of processing (crawling, indexing, rendering) preferably is completed and the index and snapshot files that result are loaded into a database or are used to update a database, maintained on the server that accepts user search criteria and composes and sends to the user the search results.

Detailed Description Text (79):
Animated GIFs and other changing features can also be identified by an icon indicating the presence of that feature. Preferably these animated features are selectively processed to provide a static image. Animated GIFs and some other technologies such as Macromedia Flash, provide an action sequence in the form of a plurality of images that are displayed in quick succession, normally in a loop. It is a problem with animations, especially those pertaining to Macromedia Flash Technology to select which frame will be captured or selected as representative of the animation. Animated GIFs begin with a graphic and the subsequent "frames" may be limited only to those pixels that have changed color from one frame to the next. Flash Technology usually begins with a blank screen or blank square. Choosing the first frame of a Flash movie as the designated frame to process and render would certainly be unaccepteable. According to alternative solutions, the Web Agent B can employ a timer to wait a predetermined time before capturing the rendered image in a file of the type that starts as a blank or fades in. It may be a matter of luck what in particular will be present at the moment captured in the changing portion of the display. An alternative is to generate a static image as a sum or average of two or more changing frames, which may produce a smeared static image. Another alternative is to disable the Flash plug in by a suitable message to the target site when loading the page. Disabling the Flash plug may eliminate any graphic data, namely if the website operators did not provide a static HTML page as an alternative to be presented for users who are not outfitted for Flash. Often, a user without Flash is presented with a blank screen with a tiny caption at the bottom reading "If you do not have Flash, click here." A rendering and subsequent snapshot of a screen similar to this could be misleading to the user if viewed within the search results of a search engine, so a timed capture is preferred.

Detailed Description Text (80):
It is an aspect of the current invention to provide an icon or similar indication within the

search results as to whether or not a particular website contains Flash Technology. This alleviates possible inconsistencies in processing and rendering a Flash movie, and subsequent interpretation by the user of a search engine who may be viewing the snapshots. Moreover, for Flash and similar technologies that are optional for users, adding an indication of their presence benefits users of the search results. Specifically in the case of Flash, a user who has loaded the Flash plugin or otherwise has the capability to process the content will prefer to access pages that contain Flash content if other factors are equal. Users with browsers incapable of processing Flash technology might be forewarned that their browser may have difficultly rendering that particular website, or at the least would be neutral about that aspect of the web site. The use of Flash, RealAudio and other "value added" technologies is often an indication that a particular website has superior content.

Detailed Description Text (93):
When the user reviews the search report using a browser, the browser inserts the graphic snapshot image adjacent to the listing of the URL link to the subject web page. Thus the user can determine whether a page entry in the search results is of interest, not only from the text information included with the URL link such as a description and title, but also from a small size presentation of what the web page looked like when it was indexed.

Detailed Description Text (95):
There are some timing issues. Between the time that the web page was downloaded and the time that the user clicks on a search result entry to review the page, the contents of the page may have changed. If a website operator updated or changed the layout of that website since it was rendered and processed by the snapshot software (Web Agent A and Web Agent B), it is possible that the visual aspect as seen through the user's browser no longer coincides with the snapshot image in the search results. Nevertheless, the snapshot normally shows a mostly consistent visual representation of the current content of the web page.

Detailed Description Text (107):
In a preferred embodiment, the textual portion of search results always is sent and caused to appear first, prior to the snapshots corresponding to those results. As a result, regardless of whether the user has turned the snapshots capability "ON" or "OFF", the text portion appears first. If a user so desires, he can abort the transmission of the results based on review of the initially received portion. This is accomplished through programming within the snapshot server system that queues the text portion of the search results to be "released" or transmitted first, preferably even before addressing (or perhaps even checking for the presence on the corresponding snapshots.

Previous Doc Next Doc Go to Doc#

L7: Entry 2 of 2 File: USPT Nov 4, 2003

6424.980

DOCUMENT-IDENTIFIER: US 6643641 B1
TITLE: Web search engine with graphic snapshots

Abstract Text (1):
A search engine manages the indexing of web page contents and accepts user selection criteria to find and report hits that meet the search criteria. The inventive search engine has an associated crawler function wherein display images of the web pages are rendered and stored as snapshots, preferably when the pages are indexed. The search engine reports search results by composing an html page with links to the corresponding page hits and containing snapshot reduced size graphic images showing the web pages as they appeared when fetched and stored as snapshots.

Brief Summary Text (12):
Search engines now operating do not search web pages on demand. Instead the search engine operators use various means to build a limited database reflecting the contents of a number of web pages. The users' search criteria are applied to the database to identify the addresses of web pages that meet the search criteria, at least from a subset of all existing web pages. Web page content can be changed. The search is current up to the most recent time at which the search engine database was updated to reflect the latest content of the web pages subject to search.

Brief Summary Text (14):
The search engine operator can use various methods to find or select web page addresses that will be loaded and analyzed or indexed in building the database. The methods may be chosen to expand or to limit the number of web pages that the search engine will access. As a result, the results of searches vary among the different search engines.

Brief Summary Text (17):
Examples of search engines include Hotbot, AltaVista, Yahoo, NorthernLight, Excite, etc. In addition, there are some search engine portals that run the same user query through a plurality of other search engines. The search engine comprises a processor that maintains a web page which the user loads by aiming his browser at the search engine URL (e.g., Excite's URL is http://www.excite.com/). The received page (namely the processed version of the html source code that is displayed) typically includes one or more Common Gateway Interface (CGI) boxes or similar form processing means by which a user who wishes to make a search enters one or more letter strings as search criteria. Boolean combinations of two or more strings often can be included or will be implied if not stated. The criteria typically are construed met if the specified words or phrases are found anywhere in the html source code of the target pages when last indexed. This includes portions that are not displayed (e.g., meta-tags and comments). The criteria can specify attributes other than the presence anywhere of a certain text string. This may be helpful, for example, to limit search results to finding files of a certain type (e.g., with URLs linking to a certain file extension type to find a certain kind of media). The criteria can also bracket out files in a selected date window.

Brief Summary Text (18):
The search engine compares the criteria to available information for web pages and sends to the user a report identifying the web pages that meet the criteria. The report to the user is transmitted in html source code. To generate the report, the search engine finds URLs for the selected web pages and inserts a list of these URLs into a shell form (i.e., an "empty" html source code file). The shell form has text and formatting to display title headers, possibly also ad banners and similar information. The URL list that is produced is inserted into the html shell. Each URL is flagged in the html source as identifying an html link (href=[etc.]).

Thus when the list is displayed by the users browser, the user can select among the results and point and click or similarly highlight and invoke the html link addressing the page that the search engine considered to meet the user's criteria. This then loads the html source code directly from the remote page that was selected and the browser displays the current contents of the referenced web page according the html source code found there at that time.

Brief Summary Text (19):
After running a search and loading the web page referenced in a URL that is mentioned by the search engine as meeting the search criteria, it is not unusual that the user may not find the loaded web page to contain the terms used as the search criteria. This occurs because the content of the page was changed to eliminate the search term between the time that it was indexed by the search engine and loaded by the user who ran the search. For the same reasons, linked pages that are reported by a search engine sometimes no longer exist.

Brief Summary Text (21):
The typical search engine reports more to the search than the URLs of the indexed pages that meet the searcher's selection criteria. The URLs themselves, which are formatted as hypertext links in the search report, sometimes provide information as to whether or not a search hit is pertinent to the user's desires. For example the domain name associated with the page may identify an owner known to be in a pertinent business, or on the contrary may show that the search result is plainly not relevant to the search. The search engine typically also stores and includes in the search report listing one or two of the first lines of the web page that is referenced, which frequently includes a title that may be helpful to show quickly whether the selected page is of interest. The search listing also may show the date at which the web page was last updated or the date that it was indexed.

Brief Summary Text (22):
The usual success rate in finding a pertinent page or website in one try or only a few tries is actually rather low. The success rate varies with the subject matter, but in a typical search the user's search criteria may turn out to be unduly broad and may select so many pages that they cannot all be reviewed, or may be so narrow that much desired content is excluded, either of which can be an unsatisfactory and perhaps frustrating experience. Balancing the needs to include relevant material and to exclude irrelevant material can result in a substantial expenditure of time, much of which is effectively wasted.

Brief Summary Text (23):
It would be advantageous if the presentation of search results could be supplemented to more effectively assist a user running a search to quickly and meaningfully separate the pertinent and irrelevant results. However, such a capability will only be useful if it can be accomplished without unduly adding processing time and storage requirements to the steps involved in preparing database information for search and in presenting the results to the user.

Brief Summary Text (25):
It is an object of the invention to provide an abbreviated representation of searchable data files, in particular Internet/Intranet/Extranet html data pages, which represents their text and linked graphics in a visual snapshot form to supplement representations such as introductory text passages and URL addresses. It is a further object to collect and process the necessary information before conducting searches and to store a relatively small graphic file in association with the search database for representing each potential hit. The respective graphics file is reported to the user when a search results in a hit on the file, namely by inserting a hyperlink to the stored file in the search report sent to the user as the search results.

Brief Summary Text (29):
These and other objects are accomplished by the improved search engine of the invention, for managing user search and selection of data files stored at distributed systems coupled at network addresses. In particular the search engine is effective to improve searching of hypertext web pages on the Internet. The search engine has an associated web crawler operable to address and load successive web pages, and to index text data associated with the successive web pages. In this manner the search engine obtains parameter information such as words appearing in documents, word proximity and other information that can be used to distinguish at least groups of the web pages from one another when conducting a search. The web crawler stores

the parameter information in a manner that cross references the paramater information with the associated web addresses or URLs of the web pages. The search engine accepts user-submitted search criteria and conducts a search or the parameter information to select the associated addresses of web pages that met all or part of the search criteria. The results can potentially be ranked, subdivided into categories and similarly handled according to known search engine operation. According to an inventive aspect, in conjunction with obtaining the parameter information for at least a subset of the web pages subject to search, the crawler renders a display image of the web page that is being indexed, and processes the image to provide a reduced size graphic image file corresponding to a static visual presentation of each of the indexed web pages. This graphic image file preferably is stored in a compressed graphic file format such as GIF, JPG, or a similar file, the file address or URL of which is stored and cross referenced to the criteria in the database that identifies the corresponding web page. When a search is conducted and results in a hit on a web page, its graphic snapshot is linked to the search results reported to the user. In a preferred embodiment, acceptance of the user search criteria and reporting of the results are handled by html page exchange communications between the search engine and the user. The search engine is accessed by the user and provides a form page having CGI boxes or the like for accepting text and/or other selections from the user. The search engine conducts a search which identifies one or more hits that are reported to the user by sending an html search results page. The search results page is composed by the search engine as a function of the search results and may contain no hits or a number of hits. Each of the hits is identified in the search results by the graphic snapshot, and preferably also by text information that reflects the content of the web page hit. Preferably, the search results page is composed to include a hypertext link to the URL address where the graphic snapshot file has been stored by the web-crawler/database/search-engine processes, for example by an IMG SRC=[path.backslash.filename] command inserted in html source code. As a result, the image file is loaded by the user's browser when processing the search results page, which generally occurs after the display of text has been accomplished.

Brief Summary Text (30):
As a result, the search results appearing on the user's browser include links to the web pages that were found to meet the criteria (hits), and also a snapshot graphic image of the way that the web page appeared when rendered at the time of indexing.

Brief Summary Text (32):
According to an inventive aspect, the graphic image file that is produced is not necessarily identical to the appearance of the page when ultimately loaded by the user after a search. In addition to the fact that the web page may have changed since it was rendered into the graphic file, the rendering is accomplished according to a predetermined display configuration of the crawler when rendered. Nevertheless, the graphic is a useful and very quick means for a user to sift through search results and determine immediately whether or not at least some of the hits bear further investigation.

Drawing Description Text (5):
FIG. 3 is a block diagram illustrating operation of the invention in connection with executing and reporting the results of searches.

Detailed Description Text (2):
According to the invention as generally shown in FIGS. 1-3, the reporting of search results by a search engine 20, is improved and facilitated by offering each searcher or user 30 a visual representation 35 of the web pages found to meet the user's search criteria submitted to the search engine. The invention is particularly applicable to an Internet search engine but can also be applied to other networks 50 where the search engine 20 is available for managing user search and selection of web pages or similar files, stored at distributed systems 52 coupled to the network. The web pages, which may be considered data files, are found at addresses to which the search engine can link to load the data files, for example being accessible using URL addressing of the pages as hypertext markup language (html), file transfer protocol (ftp), telnet or other such file types. The data files may have embedded links to other data file or to graphics or other media files. The search engine 20 of the invention accepts user queries that characterize files of interest, searches for the files and reports to each such user the results of the search including network addresses of the files found to at least partly meet the query, enabling the user to link directly to the files, and also a snapshot of how the file will appear according to the most recent rendering performed by the crawler of the search engine.

Detailed Description Text (6):
A block diagram showing an improved Internet search engine 20 according to the invention, for managing user search and selection web pages stored at distributed systems 52 coupled at network addresses to the Internet 50 or the like, is shown generally in FIG. 1. FIG. 2 illustrates a succession of method steps and/or programmed operations of the system for building and adding to or updating a database 62 of searchable information. FIG. 3 illustrates a method and apparatus for conducting searches by accepting user queries 54, conducting searches of the database 62 and reporting search results in the form of a composed search report 80 containing visual representations or snapshots 35 that depict a presentation of how the selected pages would have appeared according to a default display configuration at the time they were accessed by the crawler 60.

Detailed Description Text (8):
The search engine 20 in the embodiment shown in FIG. 1 has an associated web crawler 60 operable to address and load successive web pages from remote servers 52 on network 50, and to index or to otherwise accept or generate descriptors that characterize text data associated with the successive web pages that are loaded. In this way crawler 60 develops parameter information on the successive web pages that can distinguish at least groups of the web pages from one another, and at times can be used selectively to identify a single web page, provided some encoded aspect of that page is unique among the pages loaded and processed. The crawler 60 stores the parameter information and associated addresses of the web pages as a database 62 in a storage medium 64 that is accessible to a search processor 78 that accepts the user criteria 54 and prepares and sends search reports 80 to the query submitting user 30. The search engine portal or processor 78 responds to user submitted search criteria by searching the parameter information in the database 62 and reporting to user 30 at least the associated addresses of data files that met the search criteria when indexed. In particular, search portal/processor 78 reports the URL addresses 82 of web pages meeting the user criteria.

Detailed Description Text (14):
According to an inventive aspect, the crawler 60 that is operable to receive the web pages and to extract the parameter information from them, generates a file 72 of graphic image data corresponding to an appearance of each of the web pages, which is stored, preferably as a reduced-size and compressed image data file 75, in association with the database data respecting the page. When search results are reported to the user (FIG. 3), the search engine reports the associated URL addresses 82 of web pages that met the search criteria in a conventional manner, preferably inserting a hypertext link to each identified page into an html page reported to the user, optionally a short description or excerpt, and also inserts into the report page the graphic image snapshot file by inserting into the source of the report page a link to the stored compressed graphic image file 75. The user's browser displays the search results in conventional form, namely by showing a selectable hyperlink to the addresses and optionally a description or excerpt, and displays a snapshot of how the identified page is likely to appear if or when it is loaded by the user's browser, should the user point and click to the link to invoke the URL of the page hit.

Detailed Description Text (15):
The search portal 78 that performs the search by reference to the database 62 in storage media 64, reports the search by composing a web page containing the search results, assembling the search report using hypertext markup language. The search report contains headers and information identifying the portal and perhaps contains advertising. The search report also lists the hits that resulted from the search. More particularly, the search engine inserts (in list or table form) a text string showing the URL address of each web page hit (i.e., the pages found to meet the user criteria) together with a hypertext linkage to that URL (e.g., an "href=" statement), causing the user's browser to show a link that can be invoked (pointed and clicked) to load the page at the stated address. Such a report is conventional in an html source search report. It typically also has a description or excerpt and may be arranged in a pyramid or hierarchy of categories. According to the foregoing inventive aspect, the search engine also inserts the URL address of the graphic file that has been processed by a further process identified in FIG. 2 as Web Agent B 95, to contain a snapshot reduced/compressed graphic 35 representing the page hit.

Detailed Description Text (16):
The link to the compressed rendered graphic file can be made, for example, by use of a IMG

SRC=<domain>/<path><filename> command in the html source. The graphic can be associated with a hypertext link to the hit page URL as well as linking using an HREF=<URL of hit page> command as mentioned above. As a result, the user's browser when displaying the search results also displays the graphic snapshot image, as shown in FIG. 3.

Detailed Description Text (18):
Referring to FIG. 2., the search engine includes or is associated with web crawler 60, which is an engine that conducts web page addressing, loading and analyzing, and stores representative data in a storage device 64 containing a database 62. The stored representative data characterizes the web pages that the crawler loads and that are analyzed for content by process 68. Of the main activities to be effected by the search engine system (i.e., by the crawler and the search processor), preparation of database 62 allows a search to be conducted more quickly by reference to the processed database information gleaned from the field of possibly-selected files, than would be possible if the search engine attempted to load and analyze the entire universe of files after the user had submitted query 54 (FIG. 3), namely while the user was awaiting search results.

Detailed Description Text (22):
The database 62 is generated by preparing or obtaining a set of characterizing parameters concerning the fetched files, or their addresses or content or the like. Database 62 contains a cross reference between criteria and the identity (normally the URL address) of the file that matches the criteria. Assuming that the criteria concerns a concatenation of terms (e.g, "quick brown fox"), all the URLs of files that contain that string are available by searching for the string. Likewise the URLs of all the files containing the component terms are available ("quick" or "brown" or "fox"), and these terms or phrases can be combined with other terms or arbitrary categorizations to find a page (such as the Quick Brown Fox Hardware Store). The indexing and/or categorization particulars can be objective or arbitrary, and wholly or partly driven by human review or by automated means, and can concern any aspect that tends to be unique to individual files or common to subsets of files only.

Detailed Description Text (23):
Automated indexing and similar characterization systems may seem objective but the results are determined in part by usage chosen by the author of the content, which is to some extent arbitrary. Human review is subject to potentially arbitrary choices by the reviewer. The search database as discussed herein includes any collection of information prepared in a manner that enables search criteria to be compared to stored criteria to distinguish files from one another. The search criteria involves combinations of categorizations and/or text strings and other factors, chosen by the user in an effort to target the files or pages that have a desired subject or include reference to a particular datum. At the same time, each criterion is not applicable to every page reviewed, and as a result it is possible both to collect files that meet a user's criteria and to eliminate files that do not meet the criteria and thus are irrelevant to the particular search.

Detailed Description Text (28):
The search/reporting steps of the browser, generally shown in FIG. 3, include accepting search criteria 54 from user 30, for example using a CGI script technique in which the user enters selections including text strings, literal strings of plural terms, additional encoded aspects such as media types, date windows or limits, countries of origin, etc. The user may also select Boolean relationships (AND, OR, NOT, XOR). The search portal may require commands or may permit selection using point-and-click steps. The search engine compares the search criteria to the pre-prepared database of information gleaned from the web pages it has loaded and analyzed from the field. The results are reported to the user by preparing and formatting an html source reporting page into which hyperlinks are entered that name and point to the addresses of the files that were found to meet the criteria. Often the report includes other information such as the date the page was last updated before it was indexed, and a few lines of introductory text from the page, which provide a hint to assist the user in determining without loading the page whether the page is likely to be relevant to the search. If the user finds a link that appears to be pertinent, the user selects and engages the hyperlink. This causes the browser to load the html source found at the URL address shown in the search report, and any referenced files and links therein. However, the page may have changed between the time that the indexing was accomplished and may have totally different content than it had when indexed. The page may no longer exist. In those cases, the search fails except to advise the user that the page formerly held information that might have been of interest.

Detailed Description Text (29):
Deliberate as well as inadvertant "search engine corruption" sometimes occurs. It may be crucial for marketing or other purposes for a web site to be found in user searches on search engines, and it can be lucrative or otherwise beneficial for a web site operator if his/her site is ranked high in the search results for particular terms. Thus, a great number of website operators have ways to misrepresent the content of their pages. Keywords intended to cause the page to be selected and to rate highly in particular categories can be included and may or may not be displayed. Misleading text can be placed in miniscule font at the bottom of a page or misleading text can be hidden by making it the same color as the background on which it appears. Text can also be placed in "ALT" descriptions of images and graphics, thereby indexed by the crawler but not seen by the user. A particular term can be included one or many times to improve rankings, by one of the foregoing techniques, or by overloading keywords in "META" tags included in web pages and not displayed. Another technique is to temporarily post a page to be textually indexed by the crawler/search engine and then to replace its content after it has been indexed, or similarly, meta-refreshing the web page so as to redirect the user to another page address. According to an aspect of the present invention, the user can visually distinguish pages having undesired content and not waste time on them. Search engine corruption using the aforementioned techniques to provide misleading text is averted due to the visual nature of the present invention.

Detailed Description Text (31):
The snapshots 35 can be contained in formatted image files (e.g., GIF, JPG, etc.). The snapshot image files, or URL addresses pointing to the image files, preferably are stored in the database 62 that also contains the URL addresses of the indexed pages. In reporting search results, the search engine 78 inserts a link 82 aiming to the snapshot image file 35 into the html search results page 80. The search results appear on the users browser 84 as a link to selected pages with an associated snapshot of the page when indexed, as shown in FIG. 3.

Detailed Description Text (49):
The search engine memory also comprises text indexing data or human categorization directory data (or both), that is obtained in a conventional web crawler manner and includes an association between the text data found at each web page and the web address or URL of the originating web page. In this way, the text indexed or categorized data, and the graphic file location, are both indexed to the URL. By selecting a URL, the search engine can call up the graphic file representing its appearance when rendered at some time in the past. After receiving a selection containing one or more text strings, Boolean combinations, file extension types or other criteria, the search engine can determine the matching web pages, report their URLs and provide a graphic file showing a miniature window version of how they would have appeared if loaded by a browser at substantially the time when their data was loaded and indexed.

Detailed Description Text (51):
The search engine can comprise one or a number of processors and the processors can be in direct communication or linked on a local network or other arrangments, the key being quick access to the stored database of data representing the universe of web pages that have been processed and therefore are searchable. The search engine accepts user search criteria in a conventional way, such as using CGI form boxes to enter text strings into an associated search engine entry html page that is addressable by a browser. The search engine permits selections to be made according to at least one search criterion and preferably accepts a variety of different criteria types and combinations. These aspects of the search engine can be of the type conventionally used by current search engines such as Hotbot, Yahoo, AltaVista, Northern Light, etc. The search engine is operable to select web page hits as a function of user supplied search criteria and to determine the URL addresses of web pages (hits) that wholly or partly meet the criteria. In addition to determining the URLs of hits, the search engine may store and retrieve a brief exemplary text string such as the initial few lines of text in the web page hit.

Detailed Description Text (52):
The search engine reports search results to the user that entered the search criteria, by composing an html source page and transmitting it to the user. This html report page may identify no hits or a long list of hits, depending on the search results. In composing the report page, the search engine typically shows the search criteria used, and displays indicia

summarizing or similarly identifying each web page hit. For example, the search report can identify hits by the URL of the originating web page. Preferably a short text selection such as the first few lines of text is shown. The html coded report page prepared by the search engine includes an associated hyperlink to the URL of each hit. The URL can be shown in plain text and provided with an associated hypertext link (href=[URL]). The user reviews the URLs, sample text or other information and activates the hyperlink of a selected web page identified in the results, thereby loading the web page presently found at the address of the originating page when processed by the crawler robots.

Detailed Description Text (64):
The two general functions associated with preparing the database of information which is then subject to search and reporting, are the functions of retrieving all webpage data (performed by Web Agent A), and generating a "snapshot" file from the data (performed by Web Agent B). It is found that these functions can operate concurrently with or apart from the search engine processor or processors that search the database of information and return results to the requesting user. The preferred embodiment, however, is to perform all processing in regards to rendering, resizing, and compressing the snapshot prior to being accessible to surfers on the web. A cycle of processing (crawling, indexing, rendering) preferably is completed and the index and snapshot files that result are loaded into a database or are used to update a database, maintained on the server that accepts user search criteria and composes and sends to the user the search results.

Detailed Description Text (79):
Animated GIFs and other changing features can also be identified by an icon indicating the presence of that feature. Preferably these animated features are selectively processed to provide a static image. Animated GIFs and some other technologies such as Macromedia Flash, provide an action sequence in the form of a plurality of images that are displayed in quick succession, normally in a loop. It is a problem with animations, especially those pertaining to Macromedia Flash Technology to select which frame will be captured or selected as representative of the animation. Animated GIFs begin with a graphic and the subsequent "frames" may be limited only to those pixels that have changed color from one frame to the next. Flash Technology usually begins with a blank screen or blank square. Choosing the first frame of a Flash movie as the designated frame to process and render would certainly be unaccepteable. According to alternative solutions, the Web Agent B can employ a timer to wait a predetermined time before capturing the rendered image in a file of the type that starts as a blank or fades in. It may be a matter of luck what in particular will be present at the moment captured in the changing portion of the display. An alternative is to generate a static image as a sum or average of two or more changing frames, which may produce a smeared static image. Another alternative is to disable the Flash plug in by a suitable message to the target site when loading the page. Disabling the Flash plug may eliminate any graphic data, namely if the website operators did not provide a static HTML page as an alternative to be presented for users who are not outfitted for Flash. Often, a user without Flash is presented with a blank screen with a tiny caption at the bottom reading "If you do not have Flash, click here." A rendering and subsequent snapshot of a screen similar to this could be misleading to the user if viewed within the search results of a search engine, so a timed capture is preferred.

Detailed Description Text (80):
It is an aspect of the current invention to provide an icon or similar indication within the search results as to whether or not a particular website contains Flash Technology. This alleviates possible inconsistencies in processing and rendering a Flash movie, and subsequent interpretation by the user of a search engine who may be viewing the snapshots. Moreover, for Flash and similar technologies that are optional for users, adding an indication of their presence benefits users of the search results. Specifically in the case of Flash, a user who has loaded the Flash plugin or otherwise has the capability to process the content will prefer to access pages that contain Flash content if other factors are equal. Users with browsers incapable of processing Flash technology might be forewarned that their browser may have difficultly rendering that particular website, or at the least would be neutral about that aspect of the web site. The use of Flash, RealAudio and other "value added" technologies is often an indication that a particular website has superior content.

Detailed Description Text (91):
Upon completion of a full crawl, rendering of each and every desired web site, and full data storage of the resulting graphic snapshots, the search engine database is ready to accept user

queries. The user presents combinations of text string expressions in a known manner. According to the same sort of search criteria known in other search engine applications (e.g., HotBot, AltaVista, Yahoo, etc.), the criteria are compared to the indexed text information. By whatever means used (e.g., all words, any word, exact phrase, Boolean combinations, with or without results ranking or categorization, etc.) the search engine selects and prepares a list of the web page hits discovered by comparing the search criteria to the contents of the indexed database.

Detailed Description Text (93):
When the user reviews the search report using a browser, the browser inserts the graphic snapshot image adjacent to the listing of the URL link to the subject web page. Thus the user can determine whether a page entry in the search results is of interest, not only from the text information included with the URL link such as a description and title, but also from a small size presentation of what the web page looked like when it was indexed.

Detailed Description Text (95):
There are some timing issues. Between the time that the web page was downloaded and the time that the user clicks on a search result entry to review the page, the contents of the page may have changed. If a website operator updated or changed the layout of that website since it was rendered and processed by the snapshot software (Web Agent A and Web Agent B), it is possible that the visual aspect as seen through the user's browser no longer coincides with the snapshot image in the search results. Nevertheless, the snapshot normally shows a mostly consistent visual representation of the current content of the web page.

Detailed Description Text (107):
In a preferred embodiment, the textual portion of search results always is sent and caused to appear first, prior to the snapshots corresponding to those results. As a result, regardless of whether the user has turned the snapshots capability "ON" or "OFF", the text portion appears first. If a user so desires, he can abort the transmission of the results based on review of the initially received portion. This is accomplished through programming within the snapshot server system that queues the text portion of the search results to be "released" or transmitted first, preferably even before addressing (or perhaps even checking for the presence on the corresponding snapshots.

CLAIMS:

19. An improved Internet search engine for managing user search and selection of web pages stored at distributed systems coupled at network addresses to the Internet, the search engine having an associated web crawler operable to address and load successive web pages, and to index text data associated with said successive web pages so as to obtain parameter information that distinguishes at least groups of the web pages from one another, the crawler storing the parameter information and associated addresses of the web pages, and the search engine being operable responsive to user submitted search criteria to search the parameter information and to report at least the associated addresses of web pages that met the search criteria when indexed, wherein the improvement comprises: said crawler being operable in conjunction with obtaining the parameter information for at least a subset of said successive web pages to generate a graphic image file containing a visual image that is substantially identical to an appearance of said web pages, for display in a size proportionally smaller than said web pages; and wherein the search engine is operable when reporting the associated addresses of web pages that met the search criteria to include a representation of the graphic image file in said proportionally smaller size.

22. The improved Internet search engine of claim 21, wherein the search engine reports to the user the associated addresses of the web pages that met the search criteria, in a form of hypertext source data containing URL links to said web pages, and wherein the graphic image file is displayed in association with a URL link to the web page represented by the graphic image file.

Previous Doc Next Doc Go to Doc#